Conference: The 3rd symposium in data mining applications, (SDMA2014), Prince Sultan University, *Riyadh, KSA, 8th of May 2014.*

Using the Data Mining Techniques for Breast Cancer Early Prediction

Samar Al-Qarzaie, Sara Al-Odhaibi, Bedoor Al-Saeed, and Dr. Mohammed Al-Hagery*
*Department of IT, College of Computer, Qassim University; Saudi Arabic

***Abstract*: Data mining (the analysis step of the "knowledge Discovery in Databases" process, or KDD), a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets. The main objective of this paper is using the data mining techniques and tools for breast cancer early prediction, and to connect the technical field with medical field to serve the community. The research problem is that the specialists are not using the accumulated medical data for prediction purpose. This problem caused the loss of time and effort in hospitals and spending lots of efforts and costs and repeating the same treatments on the same cases. In this paper our focus is to use data mining techniques and tools to extract knowledge from medical datasets which related to breast cancer disease. This dataset collected from the international Repository of Artificial Intelligence. Furthermore, the research used that dataset and applied it in data mining Tools and algorithms. The data mining technique and tool used was the Decision Tree technique and WEKA tool.**

Index Terms: Data Mining, KDD processes, Decision Tree, WEKA, and medical datasets, Knowledge discovery.

I. INTRODUCTION

Currently, data mining techniques are employed to extract knowledge from medical datasets which related to breast cancer disease. Where the intelligent method applied in an essential computer-assisted process to dig through and analyzing an enormous sets of data and then extract specific data patterns in assist of tools to predict behavior along with the future trends allowing the beneficiary to make a proactive, knowledge- driven decision and to answer question that were traditionally too time consuming and expensive to resolve [1].

The overall goal of the data mining process is to extract advanced information from a data set and transform it into an understandable structure for further use [2]. In data mining the strengths and weaknesses of each of the new methods can be demonstrated on data by discussing applications of these techniques to current problems in all domains. Data mining can help in the area of medicine and pharmaceuticals to better determine which patients benefit from a given treatment [1]. The data mining provides information about cancer, including state of the art information on cancer screening, prevention, treatment and supportive care, and summaries of clinical trials [3]. When creating a data mining model, must first specify the mining function then choose an appropriate algorithm to implement the function if one is not provided by default [4].

Several works were done in this field especially in Breast Cancer Diagnosis by other methods such as using k-Nearest Neighbor with Different Distances and Classification Rules as in [6]. Other methods are used the Classification Based on Associations rules [5].

The research problem is founding a huge amount of medical datasets are not applied to assists the specialists and professionals in the medical field. The main objective of this paper is to use data mining techniques to extract knowledge from medical datasets.

The rest of the paper is organized as follows. Section 2 presents the technique and tool; section 3 shows the analysis of result followed by conclusions in section 4.

II. PROPOSED TECHNIQUE AND TOOL

The proposed tools and techniques include WEKA tool and Decision Tree technique.

Data mining tools are software components. We will use the WEKA software tool because it is support several standard data mining tasks. WEKA is a collection of machine learning algorithms for solving real-world data mining problems. It is written in Java and runs on almost any platform.

The proposed technique that will be applied in this project is Decision Tree (C4.5) because it is powerful classification algorithms that are becoming increasingly more popular with the growth of data mining.

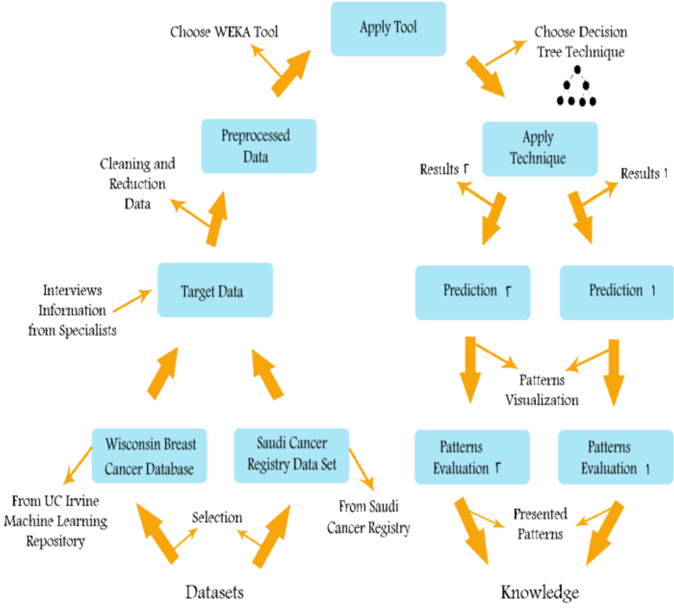
Choose WEKA Tool
Apply Tool
Choose Decision Tree Technique
Preprocessed Data
Apply Technique
Cleaning and Reduction Data
Results ٢
Results ١
Interviews Information from Specialists
Target Data
Prediction ٢
Prediction ١
Patterns Visualization
Wisconsin Breast Cancer Database
Saudi Cancer Registry Data Set
Patterns Evaluation ٢
Patterns Evaluation ١
From UC Irvine Machine Learning Repository
Selection
From Saudi Cancer Registry
Presented Patterns
Datasets
Knowledge

Figure 1: General structure of the research

Also, the Wisconsin Breast Cancer Database was used. It was collected by Dr. William H. Wolberg, University of Wisconsin Hospitals, and Madison. It has 10 attributes plus the class attribute and 699 instances. The research methodology consists of several steps and sub steps, including the data collection and data analysis and data preprocessing and analysis. The final step is to generate the results, and interprets the final results, as illustrated in figure 1.

III. RESULTS AND EVALUATION

We separated the dataset that contain 699 instances into two sets are training and testing. The training set contains 500 instances otherwise the testing set contains 199 instances. The results that were concluded are as follows:

The error rate was 6.532% and the accuracy was 93.467% for the whole results, kindly see figure 2.

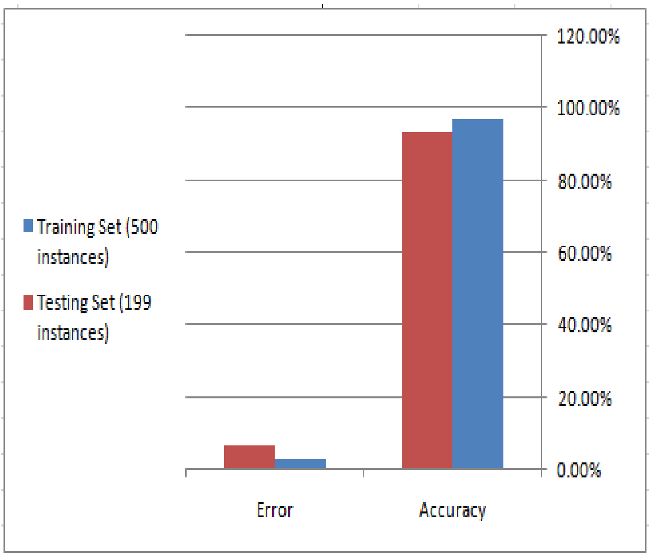

Figure 2: The results obtained from the training and testing sets.

This means that the generated results were more accurate. Thus this is best manner to test predict by WEKA program. Figure 1, shows the results obtained from the training and testing sets.

IV. CONCLUSIONS AND FUTURE WORK

This paper presents an approach for using data mining techniques (decision tree) for the early prediction of breast cancer disease. An important challenge in data mining and machine learning areas is to build precise and computationally efficient classifiers for Medical applications. Our results show that this model is very strong because the proportion of correct in testing set is 93.467% and the proportion of error is 6.532%. Whereas the proportion of correct in training set is 96.8% and the proportion of error is 3.2%. We have achieved in this paper connect the technical field with medical field to serve the community. In addition to that, this work provided practical experiences to reduce time and effort and to assists physicians to predict the disease earlier. In the future, this research work can be intended to find another data sources to use it in the test and prediction. Also, apply other types of algorithms in one time to choose the best way to get the highest accepted values based on updated data sets.

REFERENCES

[1] Zhu L, Wu B, Cao C., "Introduction to medical data mining", College of Automation, Chongqing University, 2003 Sep, http://www.ncbi.nlm.nih.gov/pubmed/14565039

[2] Mawuna Remarque KOUTONIN, "The Best Data Mining Tools You Can Use for Free in Your Company", March 8th, 2013, http://www.siliconafrica.com/the-best-data-minning-tools-you-can-use-for-free-

[3] Andrea L. Houston1, Hsinchun Chen1, Susan M. Hubbard2, Bruce R. Schatz3, Tobun D. Ng1, Robin R. Sewell4 And Kristin M. Tolle, "Medical Data Mining on the Internet: Research on a Cancer Information System", Kluwer Academic Publishers, 2000, http://ai.arizona.edu/intranet/papers/Medical-99.pdf

[4] Kathy L. Taylor, "Oracle Data Mining Concepts", Oracle and/or its affiliates, June 2013, http://docs.oracle.com/cd/E11882_01/datamine.112/e16808.pdf#page22.

[5] M., Bharati M. Ramageri, "DATA MINING TECHNIQUES AND APPLICATIONS", Indian Journal of Computer Science and Engineering, 2012.

[6] Seyyid Ahmed Medjahed, Tamazouzt Ait Saadi, "Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules", International Journal of Computer Applications, Volume 62, No. 1, January 2013.